The 'K' in 'Semantic Web' Stands for 'Knowledge'

Beek, W.G.J.

2018

document version
Publisher's PDF, also known as Version of record

SUMMARY

With the invention of the Internet and the Semantic Web, the material conditions for building an automated system that can store the world's knowledge has arisen. While pre-Internet KR systems were relatively small and homogeneous, Linked Data is produced and consumed by a large number of people and organizations who interchange knowledge on a global scale using open standards.

In this thesis, we claim the these differences in the material conditions of the production and interchange of knowledge on the Semantic Web has profound implications for the way in which Knowledge Representation and Reasoning (KR) research can be and – we argue – should be conducted. Specifically, we identify a set of simplifying assumptions under which KR research in the past was conducted, which no longer hold true under Big Data KR. This development is also observed in the Semantic Web field, where research on Description Logics and the study of relatively small ontologies has gradually been extended to cover increasingly larger collections of Linked Open Data (LOD). Nevertheless, even in most Semantic Web research of the past decade, many of the simplifying assumptions are still present.

QUALITY Even though Linked Data is intended to be interpreted by machine agents, *data quality* is often too low to make this work in practice. We show that it is possible to analyze, assess and improve the quality of the LOD Cloud as a whole by fully automated means. By automatically cleaning hundreds of thousands of Linked Data documents in LOD Laundromat, we are able to analyze and improve data quality on a large scale.

ACCESS While Linked Data is standardized by the web community, data is still disseminated in idiosyncratic ways or is limited by APIs that do not allow data to be accessed in a scalable way. With Linked Data-as-a-Service we show that it is possible to facilitate query access to (a very large subset of) the LOD Cloud for thousands of users, while running on university hardware.

HOMOGENEITY Even though the Semantic Web consists of hundreds of thousands of datasets, contemporary Semantic Web research evaluations are conducted over two datasets on average. Since Linked Datasets vary greatly in terms of size, structure, and level of semantic detail, it is unclear to what extent evaluation results obtained over a small number of datasets can be translated to the Semantic Web as a whole. With LOD Lab we show that it is possible to take the true heterogeneity of the Semantic Web into account during research evaluations.

CONTENT-INDEPENDENCE The standardized semantics for Linked Data assigns meanings to assertions in a context-independent way. This allows information to be interchanged across contexts. At the same time, assertions on the web are known to be true in some contexts but not in others. An instance of this is the use of the `owl:sameAs` predicate, which denotes the identity relation. We show that even though the official semantics of identity leads to incorrect results when applied to the Semantic Web, it is still possible to assign a context-dependent meaning in such cases, and that the context-dependent meaning can be assigned by automated means.

DECLARATIVENESS The Semantic Web characterizes the meaning of (collections of) declarative statements in terms of the model-theoretic interpretations under which those statements are true. The architects of the Semantic Web already recognized that there are other aspects of meaning, sometimes referred to as 'Social Meaning', that cannot be captured by such a declarative semantics. We focus on one particular instance of non-declarative meaning: naming. Since we cannot currently quantify the amount of non-formal meaning that is not also encoded formally, we use declarative meaning proxies to quantify the overlap between non-formal and formal meaning. We show that for most datasets there is a significant overlap between the formal meaning that is encoded in the declarative meaning proxies and the non-formal meaning that is encoded in names. As such, our approach gives a lower bound for the amount of non-declarative meaning that is encoded in the Semantic Web.

Logic and KR have been analytic research fields: they have studied how meaning should be properly represented and reasoned over. We hope that the availability of the Big Data KR infrastructures and approaches that are presented in this thesis will enable empirical studies of meaning in addition to existing analytic approaches. Such an 'empirical turn' in logic and KR will allow us to learn how meaning is actually being represented and used (correctly or incorrectly) in practice.